

Spatial-temporal association between fine particulate matter and daily mortality ¹

Jungsoon Choi, Montserrat Fuentes, and Brian J. Reich

Abstract

Fine particulate matter (PM_{2.5}) is a mixture of pollutants that has been linked to serious health problems, including premature mortality. Since the chemical composition of PM_{2.5} varies across space and time, the association between PM_{2.5} and mortality could also change with space and season. In this work we develop and implement a statistical multi-stage Bayesian framework that provides a very broad, flexible approach to studying the spatiotemporal associations between mortality and population exposure to daily PM_{2.5} mass, while accounting for different sources of uncertainty. In Stage 1, we map ambient PM_{2.5} air concentrations using all available monitoring data (IMPROVE and FRM) and an air quality model (CMAQ) at different spatial and temporal scales. In stage 2, we examine the spatial temporal relationships between the health end-points and the exposures to PM_{2.5} by introducing a spatial-temporal generalized Poisson regression model. We adjust for time-varying confounders, such as seasonal trends. A common seasonal trends model is to use a fixed number of basis functions to account for these confounders, but the results can be sensitive to the number of basis functions. In this study, the number of the basis functions is treated as an unknown parameter in our Bayesian model and we use a space-time stochastic search variable selection approach. We apply our methods to a data set in North Carolina for the year 2001.

Key words: air pollution; Bayesian hierarchical models; conditional autoregressive models; computer models; spatial epidemiology.

¹J. Choi is a graduate student at the Department of Statistics in North Carolina State University (NCSU), Raleigh, NC 27695-8203. M. Fuentes is an associate professor at the Department of Statistics in NCSU. (Email: fuentes@ncsu.edu), and B. J. Reich is a postdoctoral fellow at the Department of Statistics in NCSU. The research conducted by Fuentes and Choi has been partly supported by the National Science Foundation grants DMS 0353029 and DMS 0706731, and Reich's research has been supported by the National Science Foundation grant DMS 0354189.

1 Introduction

Spatiotemporal analyses have generated core epidemiologic data and provided important scientific basis for the recently tightened $\text{PM}_{2.5}$ air quality standard. Over the last decade, multi-city time-series studies have shown consistent associations of increased cardiopulmonary mortality and morbidity with short-term elevations of ambient $\text{PM}_{2.5}$. Some of the recent epidemiologic studies suggest that exposures to PM may result in tens of thousands of excess deaths per year, and many more cases of illness among the U.S. population (e.g., Bates, Baker-Anderson, and Sizto 1990; Dockery, Schwartz, and Spengler, 1992, Ostro et al., 1991; Schwartz 1994; Pope, Dockery, and Schwartz, 1995a, American Thoracic Society and Bascom 1996a, 1996b). However, the work by Smith et al. (2000) on fine particles, $\text{PM}_{2.5}$ ($< 2.5\mu\text{m}$ in diameter), provided evidence of lack of significant association between fine PM and mortality. Suggesting that more studies are needed, since there are remaining uncertainties and methodological challenges in understanding PM-related health effects, with respect to the uncertainty of exposure measurement errors using environmental monitoring data.

Most of the previous analyses of PM health effects have been conducted in urban areas; very little is known about the rural PM-related health effects. One reason for this is that, monitoring data are not only sparse across space but also time, since most stations only measure $\text{PM}_{2.5}$ every third or sixth day. We overcome this limitation by supplementing monitoring data with atmospheric deterministic models (e.g. CMAQ). CMAQ predicts air pollution levels at any given location and time. However, these numerical models could have a significant bias that needs to be quantified. Also, numerical models provide areal pollution estimates, rather than spatial point estimates. Thus, we have a change of support problem (see eg. Gotway and Young, 2002), since monitoring data and numerical models do not have the same spatial resolution. From our previous work on fine particles, we have developed a multi-stage spatiotemporal modeling approach which allows us to address these knowledge gaps, the change of support problem, and related uncertainties in assessing fine PM concentrations and health effects.

Recently, rigorous statistical time series modelling approaches have been used to better control for potential confounders in the epidemiological analysis of mortality associated with elevated ambient air pollutant levels. Furthermore, sophisticated analytical techniques have been introduced to adjust for seasonal trends in the data, culminating in the introduction of the generalized additive models (GAM). Although temporal trends can be explicitly included in the model, non parametric local smoothing methods (LOESS) based on GAM were widely used to take into account such trends in the analysis. Dominici et al. (2002b) suggested another approach using parametric natural cubic splines in the GAM model instead of the LOESS. One of the main limitations of this type of time series modelling approach is that it is necessary to choose the time span in the LOESS smoothing process, or the degrees of freedom of the cubic splines, and the results can be very sensitive to how that is done. In our framework, we use an alternative approach which does not involve the selection of the number of basis functions or the degrees of freedom. We estimate the shape of time-varying confounders by introducing a stochastic search variable selection (SSVS) approach (George and McCulloch, 1993) in a space-time context, while characterizing the spatial association of the time-varying confounders. SSVS was originally introduced for linear regression models and has been adopted for generalized linear models (George and McCulloch, 1997), log-linear models (Ntzoufras et al., 1997), and multivariate regression models (Brown et al., 1998). Smith and Kohn (1996) use Bayesian variable selection in a nonparametric regression model. The work presented here is the first attempt to extend Smith and Kohn’s idea to model spatiotemporal data, by randomly including/excluding basis functions from the model.

The $PM_{2.5}$ chemistry changes with space and time so its association with mortality could change across space and time. Dominici et. (2002a) showed that different cities have different relative risk of mortality due to $PM_{2.5}$ exposure. Fuentes et al. (2006) smoothed the relative risk spatially. Lee and Shaddick (2007) smoothed the risk across time. This is the first study to combine these two approaches. In our framework we allow the relative risk of mortality due to exposure to $PM_{2.5}$ vary across space and time, taking into account spatial dependencies of the mortality data and the pollution data. We show using different

model performance criteria (such as DIC) that this is a better model.

In this work we introduce an innovative hierarchical framework for spatial-temporal prediction and modelling of speciated fine particulate matter ($\text{PM}_{2.5}$) integrating atmospheric numerical models with monitoring data, and we investigate the adverse health outcomes associated with population exposure to fine particulate matter (see Figure 1). We characterize geographic differences in the $\text{PM}_{2.5}$ health effects across the state of North Carolina for the year 2001. In the first stage we incorporate multi-source and multi-level information and knowledge (monitoring network [FRM, IMPROVE], meteorological data, air quality numerical model) about ambient environment into a flexible Bayesian space-time modeling framework for estimating ambient fine PM concentrations. These refined exposure indices of $\text{PM}_{2.5}$ mass (from stage 1) are incorporated in a likelihood-based version of Poisson regression models (stage 2) to estimate the relative risks and to characterize the population susceptibility for $\text{PM}_{2.5}$ associated increases in mortality. The hierarchical framework introduced here to combine different sources of spatial-temporal data, while characterizing uncertainty and bias associated to them, is adopted to obtain more reliable estimates of air pollution levels and to reduce the variability of the relative risk parameter, that explains the association between pollution and mortality. To the best of our knowledge, this is the first study to use numerical model output in studying the association between $\text{PM}_{2.5}$ and mortality. However, this framework is flexible enough that can be adopted and implemented in many other situations where we have spatial (or spatial-temporal) information from different sources. For these data, adding CMAQ data reduces the posterior standard deviation of the relative risk for $\text{PM}_{2.5}$ by as much as 50%.

This article is organized as follows. In Section 2, we describe the different sources of data used in this study. In Section 3, we present our hierarchical Bayesian framework to study the association between $\text{PM}_{2.5}$ and mortality. In Section 4 we presents the results of this study. Finally, we provide a general discussion in Section 5.

2 Data Description

In this study we use the available PM_{2.5} data in North Carolina for the year 2001. The data were provided by the U.S. Environmental Protection Agency (EPA). The first source of PM_{2.5} data has been obtained from the Federal Reference Method (FRM) monitoring network, which includes rural and urban sites and collects PM_{2.5} samples either every day, every third day, or every sixth day. The second source of information for PM_{2.5} is from the Interagency Monitoring of Protected Visual Environments (IMPROVE) network. The IMPROVE network sites are located at national parks and wilderness areas, this network collect samples either every day, every third day, or every sixth day. Figure 2 (a) presents the yearly average of total PM_{2.5} mass ($\mu\text{g}/\text{m}^3$) at the 38 FRM monitoring sites and 3 IMPROVE monitoring sites in North Carolina for the year 2001.

Another important source of PM_{2.5} over large areas can obtained from three-dimensional (3-D) regional scale air quality models such as the U.S. EPA Community Multiscale Air Quality (CMAQ) modeling system (Binkowski and Roselle, 2003; Byun and Schere, 2006). CMAQ simulations over an airshed of interest provide gridded hourly concentrations and dry/wet deposition fluxes of major air pollutants such as PM_{2.5}. In this study we use CMAQ output from the surface layer. Figure 2 (b) presents the yearly average of CMAQ's total gridded PM_{2.5} mass ($\mu\text{g}/\text{m}^3$) for the year 2001. The CMAQ resolution used in this study is 36km \times 36km, each CMAQ value represents the averaged pollution levels within each grid cell.

Several co-pollutants (e.g., O_3) are monitored (on the hourly basis) through the State and Local Air Monitoring Stations (SLAMS), National Air Monitoring Stations (NAMS), and Clean Air Status and Trends Network (CASTNET). We have access to the SLAMS/NAMS measurements (<http://www.epa.gov/oar/oaqps/qa/monprog.html>), and CASTNET and we use them to study the influence of these co-pollutants as possible causative factors of adverse health effects. We determine the co-pollutants and fine particles effects jointly.

Daily meteorological data in North Carolina have been obtained from the U.S. National

Climate Data Center. We use the following weather variables: minimum temperature ($^{\circ}\text{C}$), maximum temperature ($^{\circ}\text{C}$), dew point temperature ($^{\circ}\text{C}$), wind speed (m/s), and pressure (hPa).

We obtained daily mortality data in North Carolina from the Odum Institute at the University of North Carolina (<http://www.irss.unc.edu>). These data include daily deaths from natural and cardiovascular causes by county in North Carolina for the year 2001.

3 Statistical Models

Our Bayesian hierarchical framework has two main stages (see flowchart in Figure 1). In the first stage we model and estimate the $\text{PM}_{2.5}$ concentrations, that will be used in the health model proposed in Stage 2.

3.1 Stage 1: Model for fine particulate matter

We introduce a spatial-temporal model for $\text{PM}_{2.5}$ using both observed data and numerical model output; this is an extension of the approach presented by Fuentes and Raftery (2005) in a purely spatial setting. We do not consider FRM measurements to be the “true” values because they are measured with error. Thus, we denote the observed total $\text{PM}_{2.5}$ mass at location $\mathbf{s} \in D_1$ on day $t \in D_2$ from the FRM network by $\hat{Z}_F(\mathbf{s}, t)$, where $D_1 = \{\mathbf{s} : \mathbf{s}_1, \dots, \mathbf{s}_{N_s}\} \subset \mathbb{R}^2$ and $D_2 = \{t : 1, \dots, T\} \subset \mathbb{R}$, and it is modeled as:

$$\hat{Z}_F(\mathbf{s}, t) = Z(\mathbf{s}, t) + e_F(\mathbf{s}, t), \quad (1)$$

where $Z(\mathbf{s}, t)$ is the unobserved “true” underlying spatial-temporal process at location \mathbf{s} and at time t . The measurement error $e_F(\mathbf{s}, t) \sim N(0, \sigma_F^2)$ is assumed to be independent of the true underlying process.

We use a similar representation for the observed $\text{PM}_{2.5}$ measurements from the IMPROVE

network, which is denoted by \widehat{Z}_I . We have,

$$\widehat{Z}_I(\mathbf{s}, t) = Z(\mathbf{s}, t) + e_I(\mathbf{s}, t), \quad (2)$$

where $e_I(\mathbf{s}, t) \sim N(0, \sigma_I^2)$ is the measurement error and is assumed to be independent of the processes $Z(\mathbf{s}, t)$ and $e_F(\mathbf{s}, t)$.

Since the CMAQ values are averages over grid squares, not point measurements, we model the PM_{2.5} CMAQ values, $\widetilde{Z}(B_b, t)$, where subregions B_1, \dots, B_B cover the spatial domain B , as follows:

$$\widetilde{Z}(B_b, t) = a(B_b) + \frac{1}{|B_b|} \int_{B_b} Z(\mathbf{s}, t) d\mathbf{s} + e_N(B_b, t), \quad (3)$$

where $a(B_b)$ is the additive bias of the CMAQ output in subregion B_b and is assumed to be a polynomial function of the centroid of the subregion, \mathbf{s}_b , with a vector of coefficients, \mathbf{a}_0 . The process $e_N(B_b, t) \sim N(0, \sigma_N^2)$ accounts for the random deviation with respect to the underlying true process and is independent of $e_F(\mathbf{s}, t)$, $e_I(\mathbf{s}, t)$, and $Z(\mathbf{s}, t)$.

The true underlying process Z is modeled as a function of the weather covariates,

$$Z(\mathbf{s}, t) = \mathbf{M}^T(\mathbf{s}, t)\boldsymbol{\zeta} + e_z(\mathbf{s}, t), \quad (4)$$

where $\mathbf{M}(\mathbf{s}, t)$ is a vector of meteorological variables (minimum temperature, maximum temperature, dew point temperature, wind speed, and pressure) with a coefficient vector $\boldsymbol{\zeta}$. The weather information is obtained from weather stations, that are not necessarily at the same locations at which we have air pollution data, thus, we have a spatial misalignment problem. To deal with this problem, we add in our hierarchical framework another level, Stage 0, in which we introduce a statistical model for the weather variables and we predict these variables at the locations of interest for Stages 1 and 2. The statistical model used for these spatial-temporal processes is the same as for the PM_{2.5} in stage 1, except for not using numerical models.

In order to predict $Z(\mathbf{s}_0, t_0)$, the true PM_{2.5} value at space \mathbf{s}_0 and time t_0 , given the

data, $\widehat{Z} = (\widehat{Z}_F, \widehat{Z}_I, \widetilde{Z})$ and \mathbf{M} , we need the posterior predictive distribution of $Z(\mathbf{s}_0, t_0)$ (see Gelfand and Smith, 1990),

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) \propto \int p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \Theta_Z)p(\Theta_Z|\widehat{Z}, \mathbf{M})d\Theta_Z, \quad (5)$$

where Θ_Z is a collection of all parameters considered in the PM_{2.5} model. After simulating N_1 values from the posterior distribution of the parameters Θ_Z , the estimator for the predictive distribution is

$$p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}) = \frac{1}{N_1} \sum_{n_1=1}^{N_1} p(Z(\mathbf{s}_0, t_0)|\widehat{Z}, \mathbf{M}, \Theta_Z^{(n_1)}), \quad (6)$$

where $\Theta_Z^{(n_1)}$ is the n_1^{th} draw from the posterior distribution.

The quantities of interest are the true total PM_{2.5} averaged over a spatial domain C_j within a county j on day t denoted by $Z_j(t)$,

$$Z_j(t) = \frac{1}{|C_j|} \int_{C_j} Z(\mathbf{s}, t) d\mathbf{s}. \quad (7)$$

Although this is not available in closed-form, the estimate of $Z_j(t)$ is obtained by averaging estimates of true PM_{2.5} values at several locations randomly chosen within a county j on day t .

Spatial priors

We use uniform priors, Unif(0,5), for σ_F and σ_I . We set these priors based on the information provided by EPA (U.S. EPA, 1997, <http://vista.cira.colostate.edu/improve/>) regarding the precision of the instrumentation used in these networks. Based on analysis of other similar datasets, we impose a uniform prior, Unif(0,5), for σ_N . In the PM_{2.5} model, based on exploratory analysis and for computational convenience, we use a separable spatial-temporal covariance function for $e_z(\mathbf{s}, t)$, which has a stationary exponential covariance for space and an autocovariance of the first-order autoregressive function, AR(1), for the

temporal component,

$$\text{Cov}(e_z(\mathbf{s}, t), e_z(\mathbf{s}', t')) = \left[\sigma_z^2 \exp\left(-\frac{h_1}{\phi_z}\right) \right] \left[\frac{\psi_z^{h_2}}{1 - \psi_z^2} \right], \quad (8)$$

where $h_1 = \|\mathbf{s} - \mathbf{s}'\|$ (in km) and $h_2 = |t - t'|$ (day). We use a $N(0, 0.1)$ prior (0.1 is the precision) for ψ_z and we use uniform priors, $\text{Unif}(1, 500)$ and $\text{Unif}(0, 100)$, for ϕ_z and σ_z , respectively.

3.2 Stage 2: Environmental Health Model

There are various statistical methods for modeling mortality data in the literature (e.g. Dominici et al., 2002a). The commonly-used model to study the association between air pollution and human health outcomes is a standard Poisson regression model with an independence assumption for the counts. However, an assumption of the Poisson model is that the mean and variance of the response variable are equal for each observation. This may be too restrictive. For example, the variance of the count data can be either smaller (under-dispersion) or larger (over-dispersion) than the mean. In this case, Poisson regression models might not be reasonable.

We use a generalized Poisson regression model (Famoye, 1993; Fuentes et al., 2006) to characterize the potential over-dispersion or under-dispersion of the mortality data. Let $Y_j(t)$ be the number of natural deaths of county j for day t , for $j = 1, \dots, J$ and $t = 1, \dots, T$. We assume that $Y_j(t)$ follows a generalized Poisson distribution (GPoi) with dispersion parameter α , mean parameter $\mu_j(t)$, and $\text{Var}[Y_j(t)] = \mu_j(t)[1 + \alpha\mu_j(t)]^2$. Based on the generalized Poisson distribution for mortality, we develop a hierarchical regression model to investigate the association between different timescales of $\text{PM}_{2.5}$ and mortality across space and season.

An important issue when studying the association between ambient $\text{PM}_{2.5}$ concentrations and daily mortality counts is whether the increased mortality associated with higher $\text{PM}_{2.5}$ levels is restricted to very frail people for whom life expectancy is short even in the absence

of PM_{2.5} exposure. This possibility is called the “harvesting hypothesis” (also known as mortality displacement). We introduce a space-time model to estimate the association between PM_{2.5} and mortality that is resistant to short-term harvesting. The method is a spatial adaption of the approach by Dominici et al. (2003) in a purely temporal context, and it is based on the assumption that harvesting alone creates associations only at shorter time scales. We use a spectral approach for the log-linear regression to decompose the information about the pollution-mortality association into distinct time scales taking into account the spatial dependency structure of the mortality and pollution data, our relative risk estimates are harvesting-resistant because we exclude the short-term information that is affected by harvesting. Thus, we decompose the daily time series of PM_{2.5} estimates for county j , $Z_j(t)$, into L orthogonal different timescales components, $Z_{j1}(t), \dots, Z_{jL}(t)$, using a discrete Fourier transform method (see Appendix).

The effect of each orthogonal decomposition of the PM_{2.5} time series is allowed to vary by county and by season. The index k refers to the seasons; we set $k = 1$ for the winter season (January-March), $k = 2$ for the spring season (April-June), $k = 3$ for the summer season (July-September), and $k = 4$ for the fall season (October-December). The parameter β_{jlk} represents the effect of air pollution for county j on the timescale l and for season k ; the log relative risk (RR) parameter is defined as $\beta_{jlk} * 10^3$. We assume

$$\begin{aligned}
 Y_j(t) &\sim \text{GPoi}(\alpha, \mu_j(t)), \\
 \log(\mu_j(t)) &= \gamma_j + \sum_{l=1}^L \beta_{jlk} Z_{jl}(t) + f_j(t) + O_j(t)\gamma_o \\
 &\quad + S_1(\text{temp}_j(t), df_1) + S_2(\text{dew}_j(t), df_2) + S_3(\text{wind}_j(t), df_3).
 \end{aligned} \tag{9}$$

The function $f_j(t)$ adjusts for the seasonality of mortality, which varies with county j . In addition to the orthogonal PM_{2.5} predictions, we also consider the co-pollutant $O_j(t)$, the daily ozone for county j and day t , imputed using a similar spatial-temporal model as in Section 3.1. The S_i 's are smooth functions of the weather covariates (temperature, dew point temperature, and wind speed) with the degrees of freedom (df) per year (df_i 's).

These weather variables are important covariates to explain air pollution.

Confounders

We consider the following confounders: age, gender, race, and hispanic/non-hispanic. Each confounder is treated as a categorical variable in our health model. We study the potential impact of these confounders on the RR by allowing an interaction between our estimated PM_{2.5} component and the different confounders. In this study the groups for each confounder are:

- Age: 0 – 14 years old (children), 15 – 64 (adults), ≥ 65 (senior adults).
- Gender: male, female.
- Race: white, black, American Indian, Other.
- Hispanic: Non-hispanic, hispanic.

Spatial priors

Since the number of deaths for each county may depend on its population size, we assume that the intercept parameter γ_j is a spatial random effect representing the baseline log relative risk of mortality for each county j . We use a conditional autoregressive (CAR) prior (Besag et al., 1991) for $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_J)^T$,

$$\boldsymbol{\gamma} \sim N(\boldsymbol{\mu}_\gamma, \sigma_\gamma^2(\mathbf{B}_+ - \rho\mathbf{B})^{-1}), \quad (10)$$

where σ_γ^2 is the overall variance parameter and ρ is the spatial association parameter. The matrix $\mathbf{B} = (B_{jj'})$ includes the neighboring information, where $B_{jj'} = 1$ if county j is adjacent to county j' , and $B_{jj'} = 0$ otherwise. The matrix \mathbf{B}_+ is a $J \times J$ diagonal matrix with elements $m_j = \sum_{j'} B_{jj'}$, $j = 1, \dots, J$. Thus, m_j is the number of “neighbors” (adjacent counties) of county j . The mean parameter $\boldsymbol{\mu}_\gamma$ has a normal prior, $N(0, 0.01)$ (0.01 is the precision). The parameter σ_γ^2 has an inverse gamma prior, $IG(0.5, 0.0005)$, as recommended by Kelsall and Wakefield (1999), the parameter ρ has a uniform prior with bounds which

are determined in order to guarantee that the variance matrix of γ is symmetric positive definite (Banerjee et al., 2004).

To account for the spatial and temporal similarity of the effect of PM_{2.5} for each timescale l , we use the multivariate intrinsic autoregressive (MIAR) prior (Gelfand and Vounatsou, 2002) for $\beta_l = (\beta_{1l}, \dots, \beta_{nl})^T$, with $\beta_{jl} = (\beta_{jl1}, \dots, \beta_{jl4})^T$:

$$\beta_{jl} | \beta_{j'l} \quad j' \neq j, \sim N\left(\frac{1}{m_j} \sum_{j' \neq j} B_{jj'} \beta_{j'l}, \frac{1}{m_j} \Sigma_{\beta_l}\right). \quad (11)$$

where the positive definite 4×4 matrix Σ_{β_l} accounts for the conditional variability as well as cross-covariance relationships between the different seasons given the neighboring sites for each time scale l . Even though the MIAR is improper, the posterior will be proper under some regulatory conditions (see e.g., Sun et al., 1999).

Seasonality of mortality

Selecting the number of basis functions to adjust for the seasonal trend of mortality is always problematic. Here, we propose an approach that avoids fixing the number of basis functions. We write the seasonal trend for county j , $f_j(t)$, using a Fourier basis (same for all counties), $C_q(t)$, $q = 1, \dots, Q$,

$$f_j(t) = \sum_{q=1}^Q c_{jq} C_q(t), \quad (12)$$

where Q is the number of basis functions and the c_{jq} 's are unknown regression parameters that control the shape of the seasonal trend at each county j . Instead of selecting the number of basis functions, we assume that Q is large enough to capture the true model and we use a Bayesian variable selection technique to stochastically include/exclude terms from the seasonal trend. We introduce a binary variable, w_{jq} , and a continuous spatial variable, r_{jq} , and express c_{jq} as

$$\begin{aligned} c_{jq} | w_{jq}, r_{jq} &= w_{jq} r_{jq}, \\ w_{jq} &\sim \text{Bernoulli}(0.5), \end{aligned}$$

where the vectors of coefficients $\mathbf{r}_q = (r_{1q}, \dots, r_{Jq})$, for $q = 1, \dots, Q$, follow independent CAR priors. If $w_{jq} = 0$, then $c_{jq} = 0$, and the corresponding basis function is not included in the model. If $w_{jq} = 1$, then $c_{jq} = r_{jq}$, and c_{jq} is non-zero. We summarize the model complexity using the posterior of $W_j = \sum_{q=1}^Q w_{jq}$, which is the number of basis functions included in the model for county j .

4 Application

We apply our statistical framework to data in North Carolina for the year 2001 to study the spatial-temporal association between daily natural and cardiovascular deaths and PM_{2.5}. We compare seasonal patterns in the effects of PM_{2.5} and its different timescales on mortality. We study the effects of ozone on mortality. Here, we decompose the daily time series of PM_{2.5} into five orthogonal components: < 3.5 days, $3.5 - 6$ days, $7 - 13$ days, $14 - 29$ days, and ≥ 30 days (Dominici et al., 2003).

The prior distribution of the spatial models in Stages 1 and 2 are described in Sections 3.1 and 3.2. In the mortality model, we use natural cubic splines for the smooth functions S_i 's with B-spline basis functions (Eilers and Marx, 1996). To select the degrees of freedom (df_i 's), we considered up to 10 df per year for each smooth function. This value seemed to be large enough based on preliminary analysis. We found that 6 df per year for temperature and 3 df per year for dew point temperature and wind speed seemed appropriate using the deviance information criterion (DIC) of Spiegelhalter et al. (2002). Since we use 1-year data, we set the number of basis functions $Q = 30$. For all MCMC sequences, we conducted a MCMC convergence diagnosis using the Gelman and Rubin (1992) convergence diagnostics, autocorrelation functions, and trace plots.

Figure 3 maps the posterior mean of the monthly average of the PM_{2.5} concentrations for January 2001 and August 2001. The estimated PM_{2.5} values in January and August were the highest in the central part of NC. Overall, the estimated PM_{2.5} concentrations in January were lower than in August. On average, the PM_{2.5} concentrations were $13.76\mu\text{g}/\text{m}^3$

for January and $16.26\mu g/m^3$ for August.

Figure 4 (a) presents the time series of the estimated $PM_{2.5}$ and its different timescales for Wake County. As expected, the plots of the short-term timescales vary rapidly from day to day, while the time series plots for the long-term timescales are fairly smooth. The $PM_{2.5}$ value for each day is the same as the value obtained by adding the values of the five timescales for that day. Figure 4 (b) shows the daily time series of mortality (total and cardiovascular disease), ozone, temperature, dew point temperature, and wind speed for Wake County. The estimated RRs at different timescales for 4 counties are presented in Figure 5. We found that the estimated RR values at longer timescale variations (≥ 14 days) are larger than those at shorter timescale variations (< 14 days) in winter and summer, with few exceptions. The standard deviation (SD) of the RR is the highest for the longest timescale (≥ 30 days), due to the potential correlation with the seasonal trend term. We also obtained estimated RR values of current day mortality using nondecomposed $PM_{2.5}$ time series. The effects of $PM_{2.5}$ on mortality in the winter and in the summer seem to be similar. The RR values of mortality by season for Wake County are summarized in Table 1. For all seasons, the RR at timescales greater than 1 month was larger than those at timescales less than 3.5 days. The effect of $PM_{2.5}$ on current day mortality in the spring was the smallest among all seasons.

We also studied the RR parameter of cardiovascular mortality by season. We found a similar pattern for all seasons, greater effects at timescales greater than 1 month than at timescales less than 3.5 days, with few exceptions. The spatial pattern of the RR for cardiovascular mortality due to $PM_{2.5}$ was similar to that of the RR for natural mortality.

We studied the impact of ozone on the association between $PM_{2.5}$ and mortality. Ozone did not seem to have a significant effect (results not shown here). For each season, the differences in the RR parameter when ozone is included in the model and when ozone is not included were small relative to the SD of the RR parameter. The 95% posterior interval was $(-0.0009, 0.0061)$.

We examine the model complexity using the estimated W_j for each county j . This index

is based on the adjustment for the seasonal trend of mortality. The posterior mean of the number of basis functions varied considerably by county. On average, the estimated number of basis functions included in the model across all counties was 10, and its SD was 2.3.

None of the confounders appeared to have a significant impact on the RR. The interaction term between the estimated $PM_{2.5}$ for the 5 timescales and the confounders was not significant across space. We conducted another study to examine the significance of the interaction term between same day $PM_{2.5}$ values and the confounders, and it was not significant either.

CMAQ

In order to examine the contribution of CMAQ to the relative risk, we repeated the analysis without the CMAQ output for $PM_{2.5}$. The posterior means of the RR parameter when the CMAQ output was not used in our model were similar to those from the full model (Figure 6 (a)). However, Figure 6 (b) shows that including the CMAQ output substantially reduces the posterior SDs of the RR. Thus, it seems that including the numerical model output improves our estimate of the effect of $PM_{2.5}$ on mortality.

Model Diagnostics and Calibration

In our generalized Poisson model, the posterior mean of the dispersion parameter α was 0.049, and the 95% posterior interval was (0.040, 0.057). This provides some evidence that the data might be overdispersed and that a generalized Poisson model is needed. We compare three different statistical models using the DIC and the root mean squared prediction error (RMSPE). The RMSPE is defined as $\sqrt{\frac{1}{N} \sum_1^N (O_i - P_i)^2}$, where O_i are the observed values at each monitoring station location, and P_i are the predicted values (using the mean of the predictive posterior distribution). We also present in parentheses the estimated effective number of parameters, p_D . The DIC for our full model was 96327 ($p_D = 1038$) and the DIC for the model with a constant RR across space was 96974 ($p_D = 1009$). The RMSPE value was also smaller for the full model (2.749) compared to the model with a RR constant across space (2.781). This justifies the need of a model that allows for spatial temporal variation in the RR, even within the relatively small geographic domain of this study. In

addition, we considered a generalized linear model (GLM) in order to assess the need for our more complex Bayesian space-time framework. We fit a traditional GLM with a Poisson model for the number of deaths and we allowed the regression coefficients to be independent over space and time, the RMSPE value of this model was 6.998. The fact that the RMSPE was almost 3 times the value obtained using our space-time model justifies the importance and relevance of taking into consideration the spatial temporal structure of the data and uncertainties associated to them.

In addition, we did calibration analysis. In Figure 7, we present at a couple of randomly selected counties (Catawba and Durham) calibration plots for the mortality analysis during the summer and the fall seasons. The percentage of the observed values that are outside the interval is 10% for the summer and 6% for the fall. Similar results were obtained at other locations. We conclude our model is well calibrated.

We conducted sensitivity analysis to study the sensitivity of the estimated RR with respect to degrees of freedom used to explain the role of the weather variables. We fit several models using 3 and 9 dfs per year for temperature and using 6 and 9 dfs per year for the dew point temperature and wind speed. When we fit each model, we used the same functions for the other weather variables. The effects at the shorter timescales were similar in all cases, while the effects at the longer timescales were slightly different. Overall, there was not a significant impact on the RR by using different dfs per year.

5 Discussion

This article presents a Bayesian framework to investigate the spatial-temporal association between $PM_{2.5}$ and daily mortality. We introduce a spatial-temporal model to obtain daily $PM_{2.5}$ concentrations by combining observed $PM_{2.5}$ data and numerical model output for $PM_{2.5}$. We estimate the association between daily mortality and different timescales of $PM_{2.5}$ to investigate the harvesting effect. Our approach to adjust for time-varying confounders does not require the selection of the number of basis functions. This hierarchical framework

takes into account the spatial and temporal dependency in the pollution and mortality data, and different sources of uncertainty about them.

The $PM_{2.5}$ and mortality association in NC is inconsistent with the harvesting-only hypothesis, and our harvesting resistant estimates of the relative risk are actually larger, not smaller, than the ordinary estimates. Our results are consistent with some other harvesting analysis (Zeger et al., 1999; Schwartz, 2000; Dominici et al., 2003). We found a similar association between different timescales and mortality for all seasons in NC. However, the association of $PM_{2.5}$ and the current day mortality in the winter is higher than in the spring in NC.

In this study, we used sparse monitoring $PM_{2.5}$ data (across space and time) as well as the CMAQ output for $PM_{2.5}$. Our results show that adding the CMAQ output reduces the amount of uncertainty in our estimated relative risk parameter.

The framework introduced here is the first step to illustrate the benefits of combining different sources of information using a hierarchical framework that allows for a space-time varying risk assessment. This approach could easily be implemented for other geographic domains, including data for the conterminous U.S. and for longer time windows.

References

- American Thoracic Society, and Bascom, R. (1996a). Health effects of outdoor air pollution. Part 1. *American Journal of Respiratory and Critical Care Medicine* **153**, 3-50.
- American Thoracic Society, and Bascom, R. (1996b). Health effects of outdoor air pollution. Part 2. *American Journal of Respiratory and Critical Care Medicine* **153**, 477-498.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall, New York.
- Bates, D. V., Baker-Anderson, M. and Sizto, R. (1990). Asthma Attack Periodicity: A Study of Hospital Emergency Visits in Vancouver, *Environmental Research*, **51**, 51-70.
- Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (With discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1-59.
- Binkowski, F.S., Roselle, S.J., (2003). Models-3 community multiscale air quality (CMAQ) model aerosol component, 1. Model description. *J. Geophys. Res.*, 108, 4183, doi:10.1029/2001JD001409.
- Brown, P. J., M. Vannucci and T. Fearn, (1998). Multivariate Bayesian selection and prediction. *J. R. Stat. Soc. B*, **60**, 627-641.
- Byun, D. W., and Schere, K. L., (2006). Review of the governing equations, computational algorithms and other components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, Applied Mechanics Reviews, 59. 51-77. ENVIRON, 2006, CAMx User's Guide, ENVIRON International Corporation, Novato, CA (www.camx.com; www.environcorp.com), September.

- Dockery, D.W., Schwartz, J. and Spengler, J. D. (1992). Air pollution and daily mortality: associations with particulates and acid aerosols. *Environmental Research* **59**, 362-373.
- Dominici, F., Daniels, M., Zeger, S. L, and Samet, J. M. (2002a). Air Pollution and Mortality: Estimating Regional and National Dose-Response Relationships *Journal of the American Statistical Association* **97**, 100-111.
- Dominici, F., McDermott, A., Zeger, S.L., Samet, J.M. (2002b) On the use of generalized additive models in time series of air pollution and health. *American Journal of Epidemiology* **156**, 193-203.
- Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. (2003). Airborne particulate matter and mortality: Timescale effects in four US cities. *American Journal of Epidemiology* **157**, 1055-1065.
- Eilers, P. and Marx, B. (1996). Flexible smoothing with bsplines and penalties. *Statistical Science* **11**, 89-121.
- Famoye, F. (1993). Restricted generalized Poisson regression model. *Communication in Statistics-Theory and Methods* **22**, 1335-1354.
- Fuentes, M. and Raftery, A. E. (2005). Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models. *Biometrics* **61**, 36-45.
- Fuentes, M., Song, H., Ghosh, S. K., Holland, D. M., and Davis, J. M. (2006). Spatial association between speciated fine particles and mortality. *Biometrics* **62**, 855-863.
- Gelfand, A. E. and Vounatsou, P. (2002). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**, 11-25.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-72.
- George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *Journal of The American Statistical Association* **88**, 881-889.
- George, E. I., and R. E. McCulloch, (1997). Approaches for Bayesian variable selection. *Stat. Sin.* **7**, 339-373.
- Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association* **97**, 632-648.
- Kelsall, J. E. and Wakefield, J. C. (1999). Discussion of "Bayesian models for spatially correlated disease and exposure data", by Best et al. In *Bayesian Statistics 6*. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), Oxford: Oxford University Press, p. 151.
- Lee, D. and Shaddick, G. (2007). Time-Varying Coefficient Models for the Analysis of Air Pollution and Health Outcome Data. *Biometrics*, doi: 10.1111/j.1541-0420.2007.00776.x.
- Ntzoufras, I., J. J. Forster and P. Dellaportas, (1997). Stochastic search variable selection for log-linear models. Technical Report. Faculty of Mathematics, Southampton University, Southampton, UK.
- Ostro, B. D., Lipsett, M. J., Wiener, M. B. and Selner, J. C. (1991). Asthmatic responses to airborne acid aerosols, *American Journal of Public Health*, **81**, 694-702.
- Pope, C. A., Dockery, D., and Schwartz, J. (1995). Review of epidemiological evidence of health effects of particulate air pollution, *Inhalation Toxicology*, **47**, 1-18.
- Schwartz, J. (1994). Air pollution and daily mortality: a review and meta analysis. *Environmental research* **64**, 36-52.
- Schwartz, J. (2000). Harvesting and long-term between exposure effects in the relationship between air pollution and mortality. *American Journal of Epidemiology* **151**, 440-448.

- Smith M, Kohn R (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–343
- Smith, R. L., Kim, Y., Fuentes, M., and Spitzner, D. (2000). Threshold dependence of mortality effects for fine and coarse particles in Phoenix, Arizona. *Journal of the Air and Waste Management Association* **50**, 1367-1379.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B* **64**, 583-639.
- Sun, D., Tsutakawa, R. K., and Speckman, P. (1999). Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika* **86**, 341-390. q
- U.S. Environmental Protection Agency (1997). National Ambient Air Quality Standards for Particulate Matter; Final Rule, Part II. *Federal Register*, 40, CFR Part 50.
- Zeger, S. L., Dominici, F., and Samet, J. (1999). Harvesting-resistant estimates of air pollution effects on mortality. *Epidemiology* **10**, 171-175.

Appendix

The daily time series of PM_{2.5} for county j , $Z_j(t)$, $t = 0, \dots, T - 1$, is decomposed into L orthogonal timescale components $Z_{j1}(t), Z_{j2}(t), \dots, Z_{jL}(t)$, where $\sum_{l=1}^L Z_{jl}(t) = Z_j(t)$. For each county j , the discrete Fourier transform is defined as

$$d_j(\omega_m) = \frac{1}{T} \sum_{t=0}^{T-1} Z_j(t) \exp(-i\omega_m t), \quad (13)$$

where $1 \leq m \leq T - 1$, i is the imaginary unit ($i^2 = -1$), and T is the length of the time series $Z_j(t)$. The m^{th} Fourier frequency is $\omega_m = 2\pi m/T$, where $0 \leq \omega_m \leq 2\pi$, and it has m cycles in the length of the data. Note that for $m \geq T/2$, $d_j(\omega_{T-m}) = \overline{d_j(\omega_m)}$, where $\overline{d_j(\omega_m)}$ is the complex conjugate of $d_j(\omega_m)$.

The inverse discrete Fourier transform is given by

$$Z_j(t) = \sum_{m=0}^{T-1} d_j(\omega_m) \exp(i\omega_m t). \quad (14)$$

Let $[0 = \omega_0, \omega_1, \dots, \omega_l, \dots, \omega_L, \pi]$ be a partition of the interval $[0, \pi]$, and we set $I_l = (\omega_{l-1}, \omega_l] \cup [\omega_{T-l}, \omega_{T-l+1})$. Then, the equation (14) is represented as

$$\begin{aligned} Z_j(t) &= \sum_{l=1}^L \left\{ \sum_{\omega_m \in I_l} d_j(\omega_m) \exp(i\omega_m t) \right\} \\ &= \sum_{l=1}^L Z_{jl}(t). \end{aligned} \quad (15)$$

Thus, $Z_j(t)$ can be decomposed into Z_{jl} 's using the following algorithm, for $l = 1, \dots, L$,

- (i) Compute the discrete Fourier transform of $Z_j(t)$ and obtain $d_j(\omega_m)$.
- (ii) Let $d_j^*(\omega_m) = d_j(\omega_m)$, if $\omega_m \in I_l$, and $d_j^*(\omega_m) = 0$, if $\omega_m \notin I_l$.
- (iii) Obtain Z_{jl} by the inverse of the discrete Fourier transform using $d_j^*(\omega_m)$, $m = 1, \dots, T/2$.

Table 1: Posterior mean (SD) of log relative rates of mortality (percent increase in mortality per increase of $10\mu\text{g}/\text{m}^3$ of $\text{PM}_{2.5}$ concentrations) for Wake County by season.

	Winter	Spring	Summer	Fall
≥ 30	18.0 (15.3)	6.5 (21.1)	33.8 (14.1)	9.9 (13.7)
14 – 29	17.8 (13.9)	0.9 (12.2)	6.9 (9.8)	-13.3 (7.7)
7 – 13	1.0 (7.4)	-2.7 (8.0)	6.3 (10.3)	10.1 (7.6)
3.5 – 6	1.6 (6.6)	1.2 (8.5)	8.0 (8.7)	-7.4 (7.6)
< 3.5	4.4 (6.5)	-3.2 (7.8)	3.3 (11.4)	-3.1 (8.2)
overall	6.5 (5.5)	0.3 (6.1)	5.1 (3.5)	3.5 (5.4)

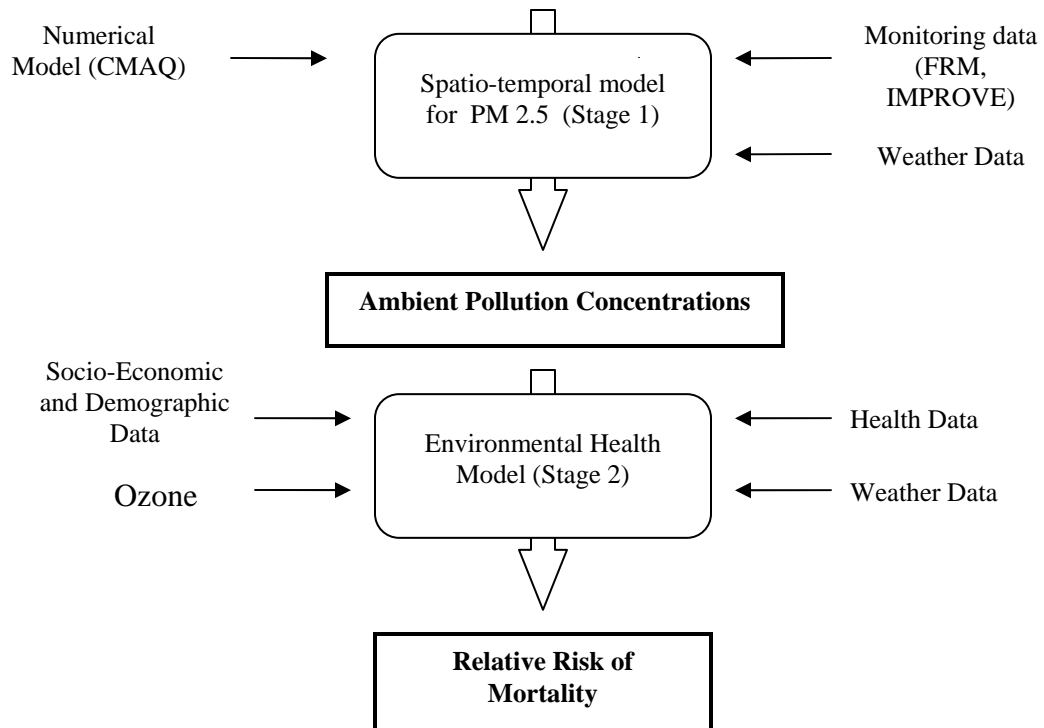
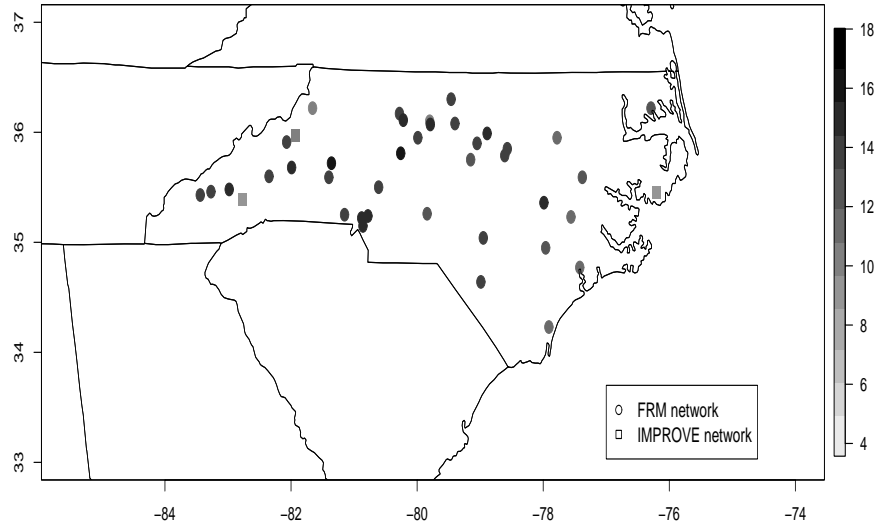
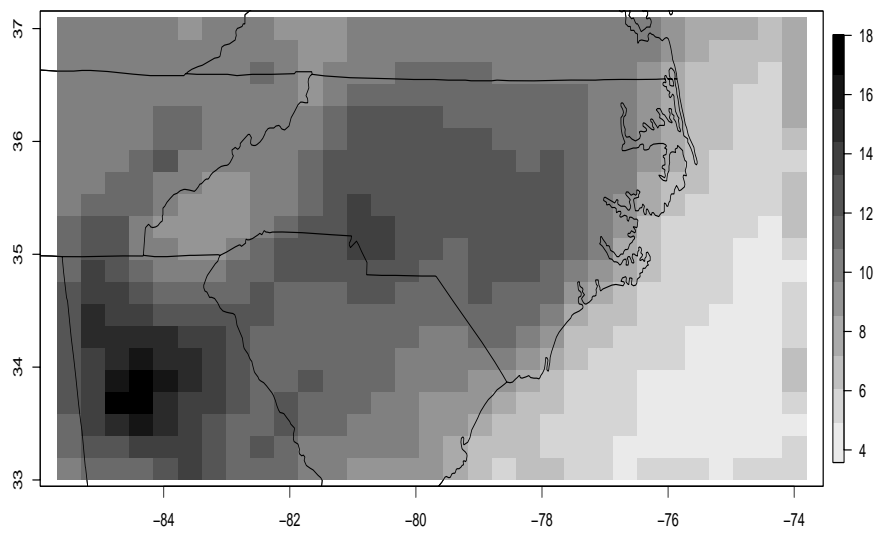


Figure 1: Hierarchical Bayesian framework to study the spatial and temporal association between fine particulate matter and mortality.

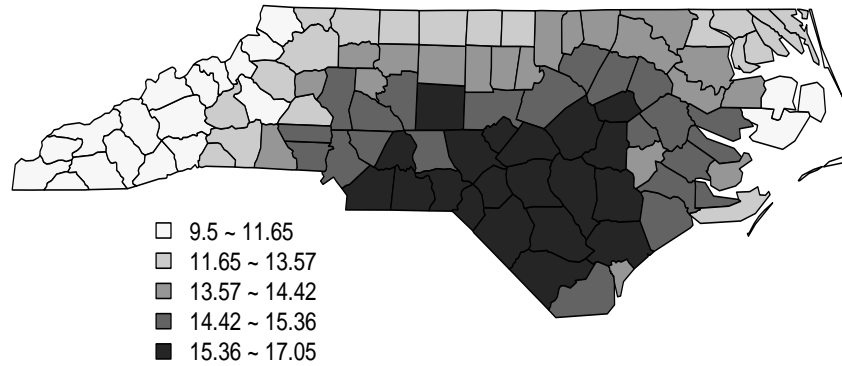


(a) FRM and IMPROVE networks

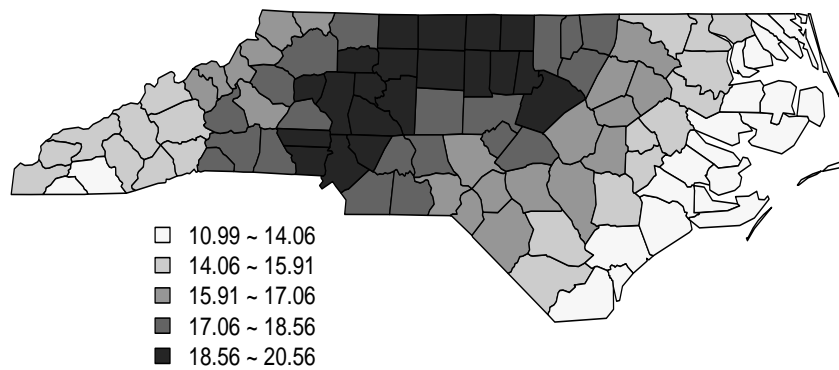


(b) CMAQ model

Figure 2: Yearly average of total $PM_{2.5}$ mass ($\mu g/m^3$) from (a) FRM network and IMPROVE network and (b) CMAQ model for 2001.

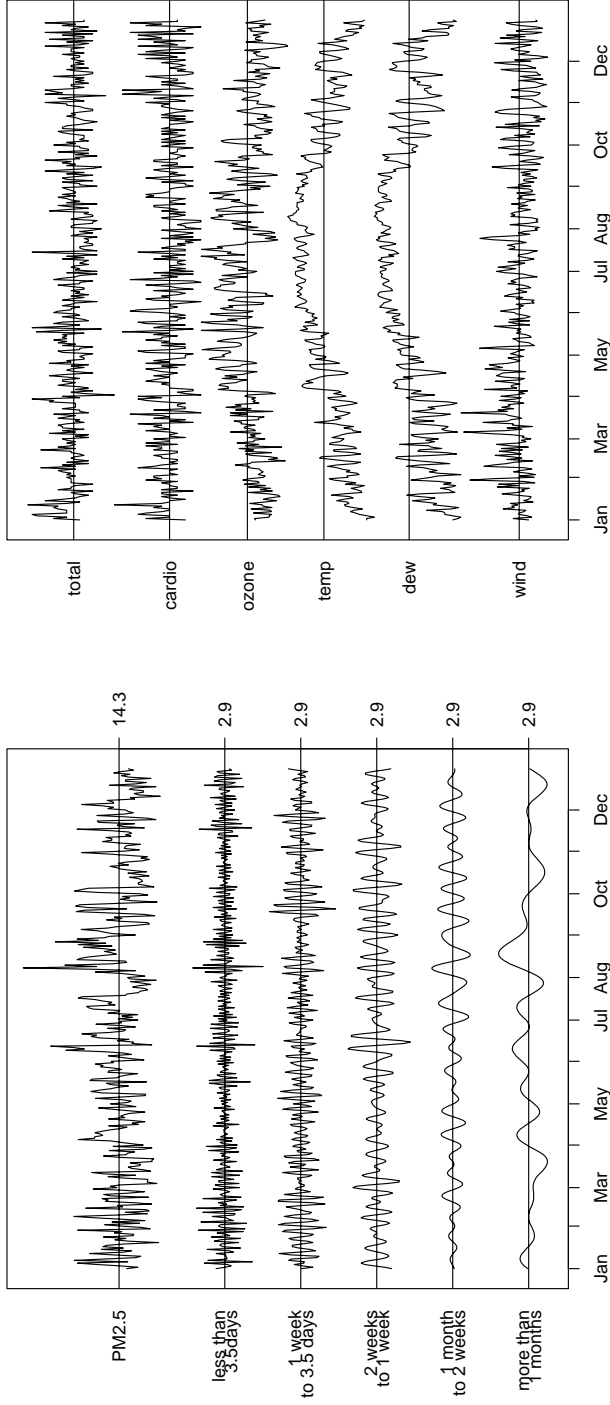


(a) January 2001



(b) August 2001

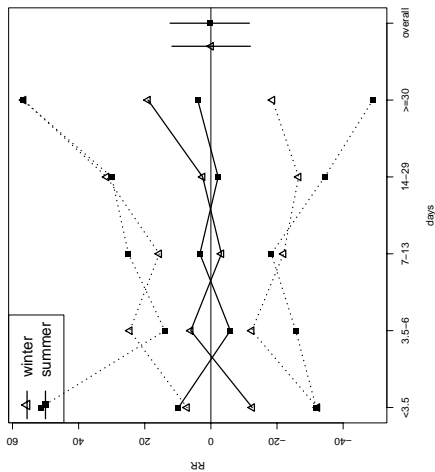
Figure 3: Maps of the monthly average of the estimated $PM_{2.5}$ concentrations for (a) January 2001 and (b) August 2001.



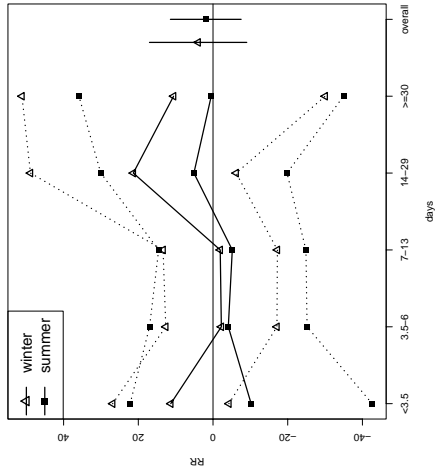
(a) Orthogonal decomposition of the PM_{2.5} time series

(b) Time series of mortality, ozone, and weather variables

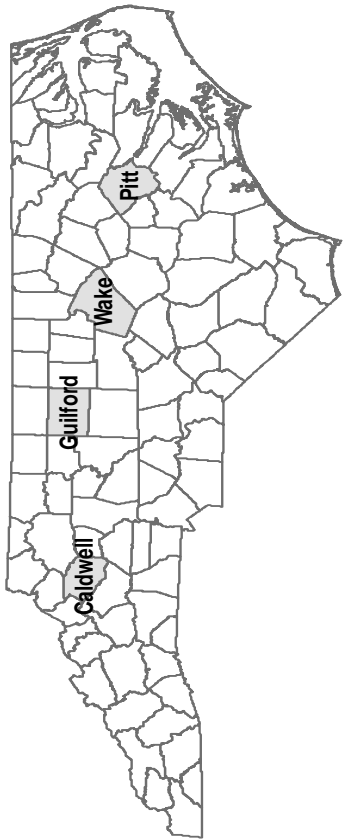
Figure 4: (a) Orthogonal decomposition of the PM_{2.5} time series and (b) time series of total natural deaths (total), cardiovascular deaths (cardio), ozone, temperature (temp), dew point (dew), and wind speed (wind) for Wake County in the year 2001. Horizontal lines show the mean value. For Wake County, the mean of the estimated PM_{2.5} is $14.3 \mu\text{g}/\text{m}^3$ and the mean of each timescale is $2.9 \mu\text{g}/\text{m}^3$.



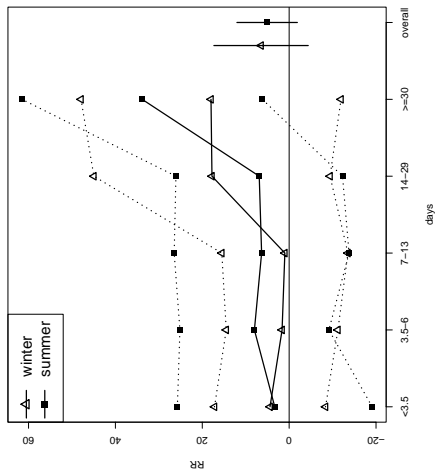
(b) Caldwell County



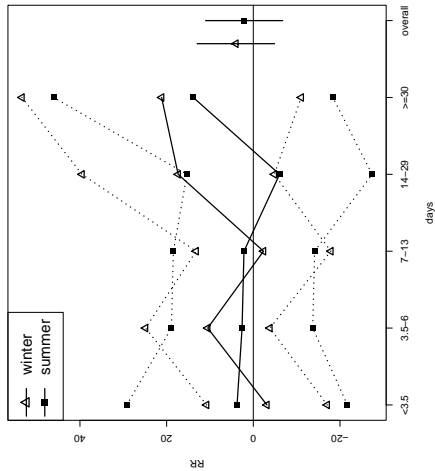
(e) Pitt County



(a) Map

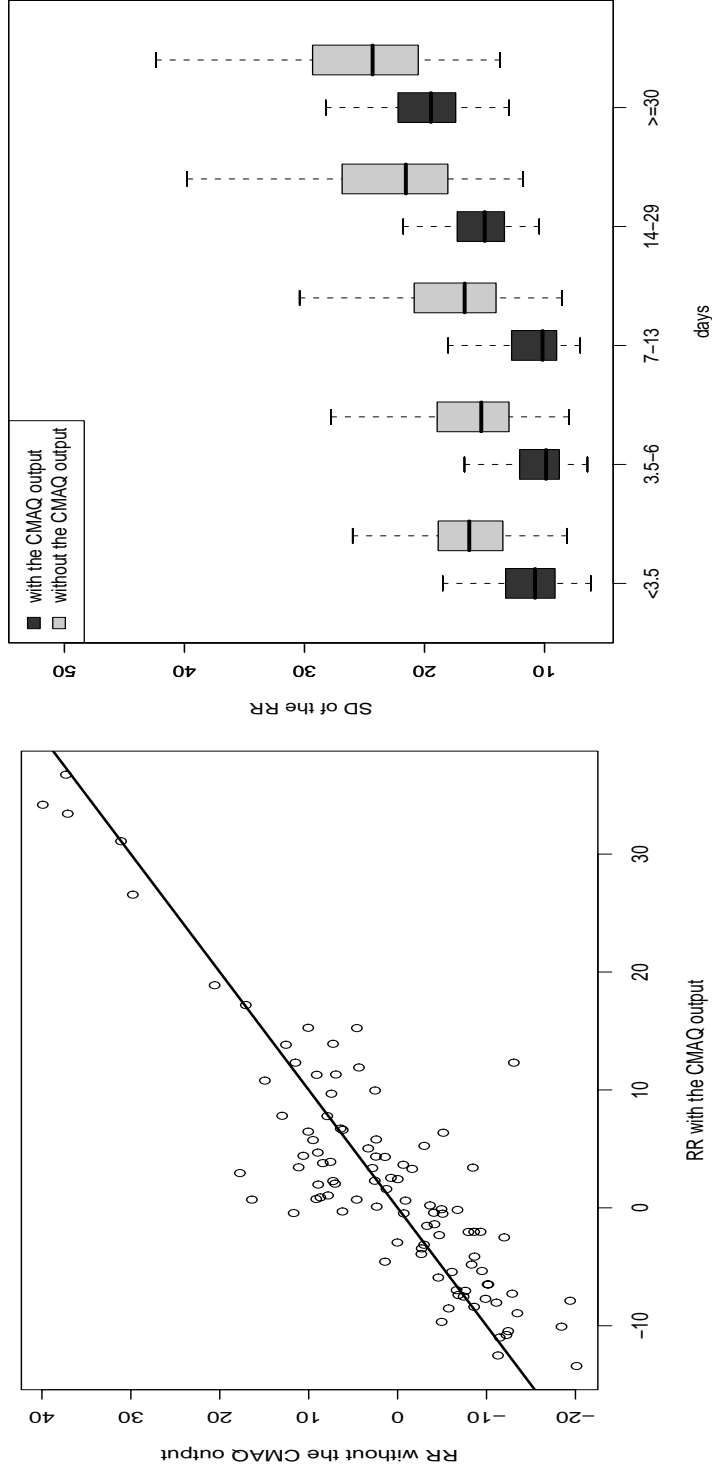


(d) Wake County



(c) Guilford County

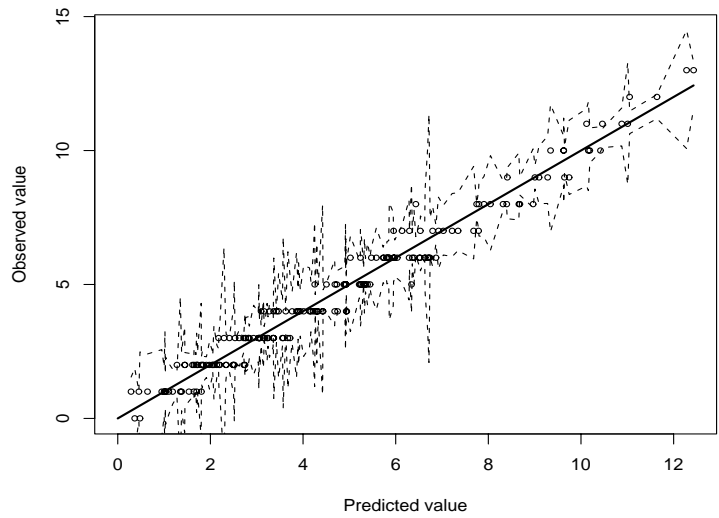
Figure 5: Map shows the location of 4 counties in NC. Mean of the posterior distribution and 95% prediction intervals for the log relative rates of mortality at different timescales (percent increase in mortality per increase of $10\mu\text{g}/\text{m}^3$ of $\text{PM}_{2.5}$ concentrations) in winter and summer. The values presented at “overall” are the estimates of log relative rates of mortality due to same-day $\text{PM}_{2.5}$ exposure.



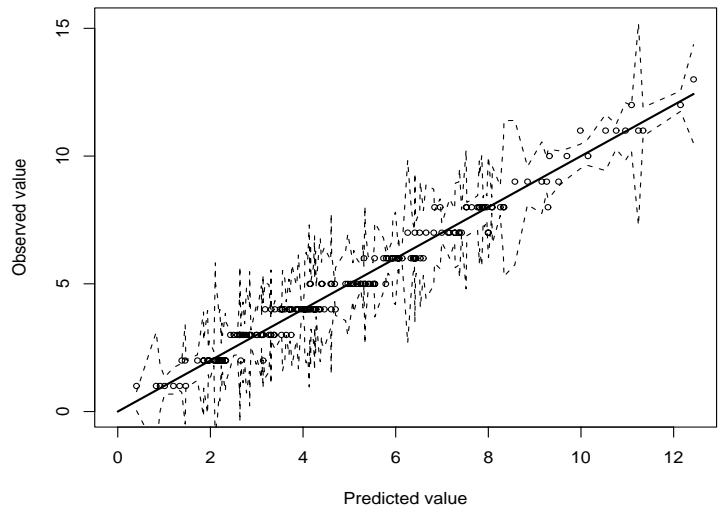
(a) Estimated RR values on the shortest timescale with and without CMAQ

(b) Estimated SDs with and without CMAQ

Figure 6: (a) Estimated RR values on the shortest timescale in the winter with and without using CMAQ output in our model and (b) Standard deviations of the estimated RR in the winter when the CMAQ output was used in the model and when the CMAQ output were not used. The solid line in (a) shows $y = x$.



(a) Summer



(b) Fall

Figure 7: Model diagnostics for mortality (a) during the summer and (b) during the fall: The dotted lines show the 95% prediction intervals.